# APART v1.0
# User Manual

## By Norma H. Pawley
## May 2004



Norma Pawley
Jason Gans
Ryszard Michalczyk

Corresponding address:
npawley@lanl.gov

## Disclaimer:

APART is provided as is. Neither the authors nor Los Alamos National Laboratory provide any warranty, or guarantee of program function or correctness of results. Individual users are responsible for the use and inferences of APART results.

# Table of Contents

# 1. Introduction

The purpose of APART (Automated Pre-processing for Assignments with Reduced Tedium) is, indeed, to remove or simplify any tedious task associated with the assignment process. Much of this tedium occurs in the preparation of peak lists for the assignment process, whether manual or automated. APART assumes that most users have existing preferences for peak-picking, and attempts to accommodate these preferences by accepting peak lists with a variety of formats. APART then refines the peak lists, using a variety of techniques that are currently often applied by hand, or with individual scripts. The goal of the refinement is to remove noise peaks from peak lists while retaining the greatest possible number of true peaks. Following peak list refinement, APART exports "clean" peak lists in the original format, and/or data formatted for the user's analysis/assignment program preference.

The flow of data through APART is illustrated in the flowchart: apart_flowchart.pdf.

## 1.1 Who is most likely to find APART useful?

- Anyone using automated peak-picking
- Anyone who is new to manual peak-picking
- Anyone planning to use AutoAssign, but not using AutoPeak
- Anyone unfamiliar with nmrView

## 1.2 Who is unlikely to find APART useful?

- Experts in peak-picking
- Experts in nmrView

# 2. License Agreement

Academic and non-profit institutions are welcome to use APART at no cost. However, users are required to send e-mail to apart@lanl.gov using the link on the program web site. The contact information provided by the user will be used for sending notification of bug fixes, significant revisions, etc.

The e-mail message should contain the following information:

User name
User e-mail address
Institution

# 3. Getting Started

## 3.1 Extracting APART

### 3.1.1 Windows

Extract using winzip

Should not need to make a compiled version:

Executable is located in the bin directory
Source code is available in the tools directory

### 3.1.2 Unix/Linux/OS X

Unzip the gzip file:

gunzip sgi_apartv1.0.tar.gz

– or –

gunzip linux_apartv1.0.tar.gz

– or –

gunzip osx_apartv1.0.tar.gz

Extract the tar file:

tar –xvf sgi_apartv1.0.tar.gz

– or –

tar -xvf linux_apartv1.0.tar.gz

– or –

tar -xvf osx_apartv1.0.tar.gz

Make the compiled version:

cd sgi_apart – or – cd linux_apart – or – cd osx_apart

make apart

## 3.2 Directory Structure

The APART home directory contains the following files and folders:

Files:

LookupTable.txt
Makefile
apart_manual.pdf – or – apart_manual.htm
apart_flowchart.pdf

Original Folders:

bin
tools

Folders created upon execution:

discards
output
statistics
working

**User input files should be placed in the APART home directory (see section 4 for file formats and minimum required input)**. With each run, data will be placed in the discards, output, statistics and working directories. Hence, after each run, the user will want to re-name these directories to avoid over-writing data.

**Examples:** In addition to the above files, sample ubiquitin data is contained in the directory ubiquitin_examples. This directory contains sample files of all the data types that APART can support.

## 3.3 Running APART

### 3.3.1 Windows

Open a command prompt (DOS) window.
Move to the APART home directory (default is windows_apart)
Invoke using

bin\apart.exe

### 3.3.2 Unix/Linux/OS X

Move to the APART home directory (defaults are sgi_apart, linux_apart, or osx_apart)
Invoke using

bin/apart

### 3.3.3 Checklist before running APART

(more detailed information provided in referenced sections):

- What is the name of my HSQC peak list?
  Does the extension reflect the appropriate file type?
  (see section 4.1)

- Do the names of my 3D peak lists conform to accepted APART file names?
  (see section 4.2)

- Do I know what my worst resolution is in each dimension?
  ($^{1}$H, $^{15}$N and $^{13}$C$\alpha/\beta$ and $^{13}$C' -- see section 5.2)

- Were my glycines collected with positive or negative intensity in the HNCACB/CBCANH ?
  (see section 5.4)

- Do I know the number of the first HSQC peak in my peak list? (see section 5.5.2)
  Note to Sparky users: your HSQC peak list is given default indexing. The number of the first HSQC peak in your peak list will be 0.

# 4. Input files

## 4.1 Accepted file types

APART will be expecting peak lists of type

.pks (AutoAssign format)
.spky (Sparky format)
.tab (nmrDraw format)
.xpk (nmrView/SmartNotebook format)
– or –
.pkl (APART format)

Examples of the expected format of these files can be found in the ubiquitin_examples directory.

**Note to Sparky users:** Sparky peak list options must be set to include peak volume information – see sample Sparky files.

### 4.2 Accepted file names

While the name of the HSQC data is flexible, all other peak files are expected to be simply named as experiment.extension. E.g., cbcaconh.xpk. The user is required to name files according to this convention. Accepted file names are discussed in more detail in section 5.

### 4.3 Minimum required input, and total accepted input

HSQC, CBCACONH and HNCACB spectra comprise the minimum required input.
In addition, HNCA, HNCACO and HNCO spectra are accepted.
HNCOCA, HCCH-TOCSY and other spectra may be implemented in future versions.

## 5. User-defined input and default values

### 5.1 Input file names

The user is prompted to enter the filename of the HSQC. The extension from this file will be used to define all other filenames. E.g., if the user enters new_hsqc.xpk, 3D spectra will be expected to be identified as cbcaconh.xpk, hnca.xpk, hncacb.xpk, hncaco.xpk and hnco.xpk. The user is required to name files according to this convention.

### 5.2 Tolerances

The user will prompted at various points to enter the expected tolerance in ppm for the $^{1}H$, $^{15}N$ and $^{13}C\alpha/\beta$ and 13C' dimensions. This number defines the maximum chemical shift difference between two peaks from different (aligned) spectra that will allow them to be defined as the same peak. The recommended tolerance is determined by the resolution of the spectra (calculated as (total sweepwidth) / (total number of points)). If spectra have been collected with different resolutions for a given dimension, the recommended tolerance is determined by the least-resolved spectrum.

When tolerances are set artificially low, peaks will be incorrectly discarded. When tolerances are set artificially high, one peak will artificially match many other peaks. The file discards/hsqc_discards.txt (Section 6.1) is a good place to look if you think your tolerances might be too low.

### 5.3 Inter-spectral Referencing

#### 5.3.1 Referencing to HSQC

APART will attempt to identify four 'remote' or 'edge' peaks in the HSQC, and will use these peaks as references for aligning all 3D spectra with the HSQC. For each 3D spectrum, the user will be provided with a default shift in each dimension (HN and 15N) that minimizes the distance between that spectrum and the HSQC. The user can over-ride this default shift if they have expert knowledge that supercedes the default alignment.

### 5.3.2 Designating $^{13}C$ reference spectrum

APART will prompt the user to identify separate reference spectra for aligning $^{13}C\alpha/\beta$ and 13C'. Default references are the CBCACONH and the HNCO, based on the expected signal/noise of these experiments. The user can choose to identify the HNCACB or the HNCACO as an alternative reference if they have expert knowledge that supercedes the default designation.

As in section 5.3.1, the 'remote' peaks will be used to define a default shift in the carbon dimension that minimizes the distance between a given spectrum and the designated reference. As in section 5.3.1, the user can override the default shift.

### 5.3.3 User satisfaction

Note that the user can shift the data as many times as desired. The program will not proceed until the user indicates satisfaction with the referencing, i.e., that no further shifting is necessary.

## 5.4 Glycine phase

APART will prompt the user to clarify whether hncacb/cbcanh spectra were collected to provide glycines with positive intensity or negative intensity. When glycines are collected with negative intensity, they can not be distinguished a priori from many beta carbons (especially: Leu, Asp, Phe, but also: Tyr, Asn, Ile), and additional input will be required from the user.

## 5.5 Interactive user editing vs. automatic filter

### 5.5.1 Overview

APART will prompt the user to accept automatic filtering, based on APART's recommendations or to proceed with interactive editing. Interactive editing is recommended as a way to further reduce the number of noise peaks in a peak list. Interactive editing can be performed by simple examination of the tabular data, **but is greatly improved when performed with simultaneous examination of spectra.** Interactive editing is a great way to identify peaks that were not picked in the initial peak lists, usually because of overlap.

### 5.5.2 Interactive Editing

Interactive editing can be very fast, or very time-consuming, depending on the quality of the spectra and the experience of the user. Progress is archived in real-time, and will be available if the user exits before completing the editing process.

If the interactive editing option is selected, the user is prompted for an output file name, or asked to accept a default (user_ab_decision.txt or user_co_decision.txt).
**Note: If a file of the same name already exists, it will be overwritten.**

The user will be prompted for the starting HSQC peak number. This is provided as a means to continue interactive editing from any point in the data, since interactive editing can be performed in multiple sessions.

If interactive editing is performed in multiple sessions, decision files should be concatenated prior to filtering.

#### 5.5.3 Filtering

Following completion of interactive editing or the decision to skip interactive editing, the program proceeds with filtering of peaks that have been identified as unlikely (noise). The user will be prompted to enter the name of the file to be filtered.

### 5.6 Compiling data

C $\alpha/\beta$ information and C' information are initially filtered and compiled separately. The user is prompted for the name of the file containing filtered C $\alpha/\beta$ information and the name of the file containing filtered C' information. The information from the two files is then woven into a single file.

### 5.7 Output file types

The user will be prompted to choose from the available output formats. These formats will be discussed further in section 6, and examples are available in the ubiquitin_examples directory.

## 6. Output files

### 6.1 discards directory

The discards directory contains three different types of discard files.

The file hsqc_discards.txt contains peaks discarded because they do not match any HSQC reference peak. This is a good place to check whether you chose your tolerances well.

Files labeled with discard_experiment.txt (e.g., discard_cbcaconh.txt) contain peaks discarded on the basis of criteria unique to that experiment (e.g., chemical shift edit, HSQC edit).

Files labeled with discard_user_*_decision.txt contain peaks discarded on the basis of information from multiple experiments (e.g., experiment-matching edit). The $8^{th}$ column from these files provides numerical references to LookupTable.txt, describing the reason a peak was categorized as noise.

### 6.2 statistics directory

The statistics directory contains files with a variety of statistics on peak volumes, tolerances for inter-spectral referencing, total applied shifts, and more. Files that are most user-relevant include track_shifts.txt, which tabulates the sum of all shifts applied to each experiment, and vol_sanity_check.txt, which tabulates max, min, and median peak volumes for each experiment.

### 6.3 working directory

The working directory contains files indicating progress at various points through the program. Helpful diagnostic files include hsqc_experiment.txt (e.g., hsqc_cbcaconh.txt), showing matches from an individual experiment to HSQC peaks, and the two sort_* files, showing matches of multiple experiments to the HSQC peaks and tentative spin system identification.

## 6.4 output directory

### 6.4.1 Files always present upon successful completion

auto_*_decision.txt:
archive of automated filtering decisions.

filt_auto_*_decision.txt:
auto_*_decision files after removal of peaks designated as noise

ab_co_together.txt
all filtered user decision data, woven together from the two filt_auto_*_decision.txt or filt_user_*_decision.txt files.

cigar.txt
matched Cα, Cβ, C' resonances. Can provide inexperienced users with potential starting points for manual or semi-automated (e.g., SmartNotebook) assignments.

no_cigar.txt
nearly matched Cα, Cβ, C' resonances. Can point to errors in spin-system identification. Potentially most useful for improving submission to programs such as Monte and Paces.

two_of_three.txt
compilation of "two out of three" matches, e.g. Cα, Cβ match (missing carbonyl) or Cβ, C' match (missing alpha). May identify potential starting points for locating peaks missing due to overlap.

session_transcript.txt
archive of all user interactions with APART not present in user_*_decision.txt.

### 6.4.2 Files dependent on user choice

user_*_decision.txt:
archive of user filtering decisions, if interactive editing was selected

filt_user_*_decision.txt:
user_*_decision files after removal of peaks designated as noise

monte.cs:
output formatted for assignment with MONTE

paces.txt:
output formatted for assignment with PACES

filt_experiment.pks (e.g., filt_cbcaconh.pks)
output formatted for assignment with AutoAssign

filt_experiment.*

(e.g., filt_cbcaconh.xpk, filt_cbcaconh.tab, filt_cbcaconh.spky)
filtered input data. Not shifted to reflect inter-spectral referencing, based on user preference (as reflected in session_transcript.txt)

filt_shift_experiment.*
(e.g., filt_shift_cbcaconh.xpk, filt_shift_cbcaconh.tab, filt_shift_cbcaconh.spky)
filtered input data. Shifted to reflect inter-spectral referencing, based on user preference (as reflected in session_transcript.txt). Useful for semi-automated assignment programs, such as SmartNotebook.

**Note that, with the exception of AutoAssign formatted data, output peak lists can only be of the same format as input peak lists.**

## 7. Trouble-shooting

APART is currently in the beta-testing stage. Hence, comments and requests for help are expected, and welcomed.

Contact: npawley@lanl.gov

The most common source of problems is always formatting. An excellent first step in the trouble-shooting process is to check file formats against the examples provided in the ubiquitin_examples directory.